CS 7545: Machine Learning Theory **Fall 2021**

Lecture 19

Name: Tyler LaBonte **Professor:** Santosh S. Vempala

1 Lower Bound on Parity

Recall that PARITY is the problem where we are trying to recover a subset $S^\star$ of Boolean variables whose parity determines the parity of the input x.

Theorem 1.1

Weak learning PARITY takes $2^{\Omega(n)}$ statistical queries of tolerance $\tau \geq \frac{1}{\text{poly}n}$.

We know how to solve PARITY when $S^\star$ is fixed, but in the SQ model, we can use distributions to make the problem difficult. Let $D = U(\{-1,1\}^n)$ and $S^\star = U([n])$. Imagine we asked the SQ oracle a parity query:

$$h(x, f(x)) = \chi_S f(x) \implies \mathbb{E}[S] = \begin{cases} 0 & S \neq S^\star, \\ 1 & S = S^\star. \end{cases} \tag{1}$$

This query will only tell us $S \neq S^\star$, but will give us no additional information, since all negative examples look the same. The SQ model relies on this additional information to eliminate larger portions of the concept space. So, if we can show that any query can be written as a parity, we will obtain the theorem.

One small technicality is that we have to extend χ_S from 2^n dimensions to 2^{n+1} to accommodate $f(x)$. We have $\chi_S(x, f(x)) = \chi_S(x)$ pairwise orthogonal, and we add $h_S(x, f(x)) = f(x)\chi_S(x)$.

Claim 1.1

χ_S and h_S together are an orthonormal basis (orthogonal and norm 1).

Proof: We clearly have

$$\langle \chi_S, \chi_T \rangle = 1 \text{ iff } S = T. \tag{2}$$

Then,

$$\langle \chi_S, h_T \rangle_D = \mathbb{E}_D[\chi_S(x)\chi_T(x)f(x)] = 0 \tag{3}$$

and

$$\langle h_S, h_T \rangle_D = \mathbb{E}_D[\chi_S(x)\chi_T(x)f^2(x)] = \langle \chi_S, chi_T \rangle = 1 \text{ iff } S = T. \tag{4}$$

■

So, any query function $g : \{-1,1\}^n \to [-1,1]$ can be written in the basis:

$$g(x, f(x)) = \sum_S \alpha_S h_S(x, f(x)) + \sum_S \hat{g}\chi_S(x), \tag{5}$$

where $\hat{g}$ is the Fourier coefficient. The SQ oracle outputs $\mathbb{E}[g] \pm \tau$. Since the second term doesn't depend on the label, we can treat it as constant. Then,

$$\begin{aligned} \mathbb{E}_D[g(x, f(x))] &= \sum_S \alpha_S \mathbb{E}_D[\chi_S(x) f(x)] + g_0 \qquad (6) \\ &= \sum_S \alpha_S \mathbb{E}_D[\chi_S(x)\chi_{S^\star}(x)] + g_0 \qquad (7) \\ &= \alpha_{S^\star} + g_0. \qquad (8) \end{aligned}$$

How many $S^\star$ can have $|\alpha_{S^\star}| \geq \tau$?

$$\sum_T \alpha_T^2 \leq 1 \text{ and } \alpha_T^2 \geq \tau^2 \implies \#S^\star \leq \frac{1}{\tau^2}. \tag{9}$$

That is, the total possible S is 2^n, but we only eliminate $1/\tau^2$ each time. So, we would need exponential queries.

We can use this framework to show SQ lower bounds on different problems included planted clique, Gaussian mixtures, and neural networks. This is particularly useful because SQ includes many algorithm classes such as EM, GD, Markov, and moments-based methods.